

APPLICATION
FOR
UNITED STATES LETTERS PATENT

TITLE: SERVING DYNAMIC WEB-PAGES
APPLICANT: MARK E. KRIEGSMAN AND BENJAMIN W. WYCKOFF

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL624269748US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

Date of Deposit September 22, 2000

Signature

Samantha Bell
Typed or Printed Name of Person Signing Certificate

SERVING DYNAMIC WEB-PAGES

The invention relates to serving dynamic web-pages, and in particular, to the updating of obsolete portions of a dynamic web-page.

BACKGROUND

5 A dynamic web-page typically includes one or more static objects and one or more dynamic objects. Serving a web-page thus includes assembling these objects and integrating them to form a web-page.

The static objects are those objects that change rarely, if at all. Examples of such objects include company logos or general instructions and forms for ordering
10 merchandise, biographies of key personnel, and general background about the company serving the web-page.

Dynamic objects are those objects that change frequently. For example, a catalog of merchandise may include notice of special sales that must be updated at regular intervals. Stock prices or news items are often updated on a daily basis. In another
15 example, a web-page can include objects that are uniquely associated with a user requesting the web-page. One common example of such an object can be found in web-pages that greet a repeat customer by name. These components must be updated every time a user requests the web-page.

A modern web-serving system is typically a distributed system in which an origin
20 server is in communication with a plurality of caches. Static components of a web-page are typically stored in the caches, whereas dynamic components of the web-page are generated by the origin server. In such a system, the process of assembling a web-page for serving to a customer involves retrieval of a variety of constituent objects from various locations. A distributed web-serving system of this type is advantageous because
25 it frees the origin server from having to provide static objects.

A disadvantage of the foregoing distributed web-serving system is that although dynamic objects need to be periodically updated, they do not always need to be updated each time a request is made. For example, a dynamic object that lists the daily price of

various commodities needs to be updated only once a day. A dynamic object that lists available items needs to be updated when the last unit of a particular item has been sold.

The updating of dynamic objects often involves execution of a database query by the origin server. For example, when a customer wishes to view what items are being offered for a special discount, the origin server may need to formulate a database query that not only asks for a list of what items are on sale but also determines the number of such items available for sale.

To further reduce the load on the origin server, it is preferable that these dynamic objects also be cached and that the cache be updated periodically to avoid the serving of obsolete dynamic components. Where the generation of a dynamic object requires execution of a database query that is expected to be common to many requesting users, it is useful to cache the results of that query to avoid having the origin server needlessly re-execute the same query.

Existing cache management software is not capable of gracefully accommodating the caching of dynamic web-pages. The conventional method of managing a cache of dynamic web-pages is to assign to each dynamic object a date on which that object becomes stale and must be updated. A disadvantage of this method is that it assumes that a stale date, which is to some extent arbitrarily assigned, is a definitive indicator of whether a dynamic object needs to be updated. In fact, there may be occasions when a dynamic object becomes obsolete long before its stale date. Conversely, there may be occasions when a dynamic object remains current long after the stale date.

SUMMARY OF THE INVENTION

A web-serving system according to the invention enables the contents of cache memory to be controlled with greater flexibility by providing a programmable script in communication with a cache manager. Because the script is programmable, it can be customized by a programmer. In particular, the programmable script can be made to detect and respond to a triggering event and to instruct the cache manager to alter the content of cache memory in response to that triggering event.

In one aspect of the invention, the web-serving system also includes a usage-monitor for collecting access-data indicative of the frequency with which a selected web-page is requested. The usage-monitor optionally provides the access-data to the programmable script. The programmable script can use this data to alter the content of cache memory in response to the access-data.

In another aspect of the invention, the web-serving system includes a communication path between the programmable script and an administrator process. This communication path enables the programmable script to receive instructions from the administrator process.

Other features and advantages of the invention will be apparent from the following detailed description, the claims, and the accompanying drawings in which:

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a distributed web-serving system incorporating the invention; and

FIG. 2 is a centralized web-serving system also incorporating the invention.

DETAILED DESCRIPTION

A system 10 embodying the invention, as shown in FIG. 1, includes an origin server 12 in communication with a cache server 14 and a database 16. The origin server 12 supports a database engine 18 for formulating and executing database queries in response to instructions issued by an administrator process 20, also supported by the origin server 12. The database 16 stores data that is used to create dynamic web-pages. For example, if the origin server 12 is associated with a retailer, the database may include current inventory and pricing information, or information regarding weekly specials. The origin server 12 is typically in communication with the cache server 14 over a first internet link 22. Although only a single cache server 14 is shown in FIG. 1, it is understood that the system 10 can include many geographically dispersed cache servers 14.

The cache server 14 supports a cache manager 24 that fetches and stores constituent objects of web-pages in a cache memory 26. These objects are provided to the

cache manager **24** by the administrator process **20** running on the origin server **12**. When necessary, the administrator process **20** provides replacement web-page objects to the cache manager **24**, which then replaces corresponding obsolete objects that are resident in the cache memory **26**.

5 A browser **28** executing on a client machine **30** typically requests web-pages from a web site served by the origin server **12**. In a distributed web-serving system **10**, the browser's requests are automatically redirected to an available cache server **14** such as that shown in FIG. 1. The browser thus establishes communication with the cache manager **24** over a second internet link **32**, as shown in FIG. 1.

10 In response to the request for a web-page, the cache manager **24** identifies the constituent objects of the web-page and where they are stored. Some of those objects are already available in the cache memory **26**. Under these circumstances, the cache manager **24** fetches these objects from the cache memory **26** and provides them to a page assembler **34** for inclusion in the web-page.

15 Other constituent objects of the web-page are not available from the cache memory **26**. For these missing objects, the cache manager **24** issues a request to the administrator process **20** on the origin server **12**. In response, the administrator process **20** prepares the missing objects, if necessary, and provides them to the cache manager **24**. The cache manager **24** provides the missing objects to the page assembler, which then
20 completes the web-page and provides the assembled web-page to the cache manager **24** for delivery to the client machine.

 The step of requesting missing objects of a web-page is a source of delay in serving clients. It is thus desirable to maintain as much of the web-page in the cache memory **26** as possible. There are two constraints that impose limits on what can be
25 stored in the cache memory **26**: the finite capacity of the cache memory **26** and the tendency of objects stored in the cache memory **26** to become obsolete.

 Because the cache memory **26** has a limited capacity, it is preferable that only the most frequently requested objects be stored in it. This requires that, at some point, a

prediction be made concerning how frequently requested a particular object is expected to be.

In some cases, this prediction is simple. If every web-page contains a graphic image with the company logo, then clearly that component should be stored in cache
5 memory 26.

In other cases, however, the popularity of a web-page may vary depending on unpredictable external events. For example, changes in fashion may gradually make web-pages offering the sale of certain items more popular. The random occurrence of newsworthy events may cause feature articles on previously obscure topics to surge in
10 popularity for a limited, and unpredictable, time.

In addition to its limited capacity, cache memory 26 has limited currency. Unless steps are taken to periodically update cache memory 26, objects stored therein will be served to clients long after they have become obsolete. In some cases, it is possible to replace objects at appropriate times because the nature of those objects renders their
15 obsolescence predictable. For example, an object listing the daily closing price of a stock, by definition, requires daily replacement.

In other cases, however, obsolescence can strike an object at any time, with little or no warning. For example, if an object lists special sale items available at a store, it is entirely possible that the sale will be so successful that inventory of certain popular items
20 will be prematurely depleted. Under these circumstances, the object may become obsolete long before it is scheduled to be updated.

Constituent objects in the cache memory 26 typically carry information specifying a date and time on which they should be replaced. In such cases, the cache manager 24 need only monitor the time and transmit a request for an updated web-page at the
25 appropriate time. This is a disadvantage because, as noted above, an object may need to be replaced prematurely, either because an unpredictable event occurred or because an event occurred at an unpredictable time. Since the conventional cache manager 24 will

only replace a web-page on its designated date, the premature replacement of a web-page is difficult.

To address the foregoing disadvantages, a system 10 according to the invention provides communication between the cache manager 24 and a programmable script 36.

- 5 The programmable script 36 can be a set of JavaScript instructions provided by a programmer. The script 36 can thus cause the cache manager 24 to update selected constituent objects of a web page upon the occurrence of a programmer-defined triggering event. One such triggering event can, of course, be the passage of a selected amount of time. However, because the script 36 is freely programmable, the triggering
- 10 event can be any event that can be defined by the script 36. For example, if more than five clients order the same item, the script 36 can issue a request for an update from the origin server 12 even though the designated replacement time for a particular object may not have arrived. The programmable script 36 thus liberates the cache manager 24 from relying solely on the passage of time as a triggering event for replacing constituent
- 15 objects of web-pages.

- If a web-page increases in popularity, or if its popularity exceeds a threshold, it may be preferable to store the constituent objects of that web-page in the cache memory 26 if they are not already there. Conversely, if a particular web-page is decreasing in popularity, or if the popularity of a web-page falls below a threshold, it may be preferable
- 20 to remove the constituent objects of that web-page from the cache memory 26 to free space for objects associated with more popular web-pages.

- To enable it to engage in such adaptive behavior, the web-serving system can further include an optional usage monitor 38 to collect access-data that illuminates the types of web-pages being requested by client machines. With the usage-monitor 38 in
- 25 place, the cache manager 24 is able to identify trends in the popularity of particular web-pages. If, on the basis of access-data provided by the usage-monitor 38, a script 36 determines that a particular web-page has become, or is about to become more popular, it can instruct the cache manager 24 to request copies of the constituent objects of that web page for storage in the cache memory 26. Conversely, if on the basis of access-data

provided by the usage-monitor 38, the script 36 determines that a web-page is no longer as popular as it once was, it can instruct the cache manager 24 to remove the constituent components of that web page from the cache memory 26.

The usage-monitor 38 and the script 36 thus combine to create a feedback control system in which the measured variables are the popularities of particular web-pages and the controlled variables are the locations of constituent objects of various web-pages. The usage-monitor 38 obtains access-data concerning the measured variables and provides them to the script 36, which functions as a controller. On the basis of that access-data, the script 36 causes the cache manager 24 to adaptively manage the contents of the cache memory 26.

The script 36 can also take access-data from the usage-monitor 38 and, instead of acting upon it as described above, it can report that access-data to the administrator process 20. This access-data can then be made available to human analysts who can then propose new rules to be implemented as scripts.

There also exist web-serving systems that are not distributed systems such as that shown in FIG. 1. Such web-serving systems also benefit from having a high-speed cache memory in which frequently requested constituent objects of web-pages are stored for rapid retrieval. A representative web-serving system of this type is shown in FIG. 2.

In the web-serving system 40 of FIG. 2, the cache memory 26, the cache manager 24, the usage monitor 38, the page assembler 34, and the script 36, all reside on the origin server 12. In the embodiment of FIG. 2, there is thus no need to redirect requests for web-pages to a cache server because there is no cache server to redirect to. Instead, requests for web-pages are directly intercepted by the usage monitor 38.

The invention can be implemented in hardware or software, or a combination of both. The invention can be implemented in computer programs using standard programming techniques following the method steps and figures described herein. The programs should be designed to execute on programmable computers each comprising a processor, a data storage system (including memory and/or storage elements), at least one

input device, and at least one output device, such as a CRT or printer. Program code is applied to input data to perform the functions described herein and generate output information. The output information is applied to one or more output devices such as a CRT, as described herein.

- 5 Each program is preferably implemented in a high level procedural or object oriented programming language to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language can be a compiled or interpreted language.

- 10 Each such computer program is preferably stored on a storage medium or device (e.g., ROM or magnetic diskette) readable by a general or special purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer to perform the procedures described herein. The system can also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to
15 operate in a specific and predefined manner to perform the functions described herein.

Having described the invention, and a preferred embodiment thereof, what we claim as new and secured by letters patent is: